The dynamics of assessment center validity: results of a seven year study

Jansen, P.G.W.; Stoop, L.A.M.

published in
Journal of Applied Psychology
2001

DOI (link to publisher)
10.1037/0021-9010.86.4.741

document version
Early version, also known as pre-print

Faculteit der Economische Wetenschappen en Bedrijfskunde

SERIE RESEARCH MEMORANDA

The Dynamics of Assessment Center Validity: Results of a Seven Year Study

Paul Jansen
Bert Stoop

Research Memorandum 2001-30

July **2001**

vrije Universiteit *amsterdam*

The Dynamics of Assessment Center Validity: Results of a Seven Year Study

Paul Jansen (Vrije Universiteit Amsterdam, The Netherlands)
Bert Stoop (KPN Research at Groningen, the Netherlands)

Abstract

We investigated tempora1 trends in the validity of an assessment center consisting of a group discussion and an analysis/presentation exercise, for predicting career advancement as measured by average salary growth over a 7-year period, for a sample of 679 academic graduates. The validity of the overall assessment rating (OAR) for persons with tenure of 7 years, corrected for initial differences in starting salaries, restriction in range, was .39. There was a considerable time variation in the validity of both the OAR and assessment center dimensions. In accordance with findings from research in managerial effectiveness and development, the dimension interpersonal effectiveness only became valid after a number of years, while the dimension firmness was predictive in the whole period and increased in time. For comparison, validity trends for two types of interviews and a mental test were also studied.

The Dynamics of Assessment Center Validity: Results of a Seven Year Study

An extensive body of knowledge exists about the predictive validity of assessment centers. Assessment centers have predictive validity for work-related criteria such as career potential or appraisal of overall job performance (Gaugler et al., 1987; Schmitt et al. 1984). However, a genera1 problem for evaluating assessment center validity, especially in the case of career advancement, is criterion contamination. In that case, later promotion decisions are indirectly and inadvertently based on the initial assessment center rating (Klimoski & Brickner, 1987). Therefore research into ***Zong-ferm,*** uncontaminated assessment center validity is needed.

But, the variation in validity of assessment centers with time is unclear (Gaugler et al., 1987). Particularly, it is unclear whether the generally observed decline in predictive validity of selection instruments (Hulin, Henry & Noon, 1990), also is the case for the assessment center (Barrett, Alexander & Doverspike, 1992). Predictive validity of assessment centers has been found to remain more or less unaffected, decrease or increase with time (Tziner, Ronen & Hacohen, 1993). In addition, there exists little research in which tempora1 trends in assessment center validity are compared to validity pattems of other predictors used in the same sample (Hunter & Hunter, 1984).

The present study investigates time-dependent pattems in assessment center validity. In addition, the incremental validity of the assessment center with respect to such other predictors as the interview and a mental test is investigated.

LONG TERM ASSESSMENT CENTER VALIDITY

Assessment centers are in particular predictive for 'advancement criteria' such as career progress, salary advancement, long-term promotion, and potential development (Bray, Campbell & Grant, 1974; Ritchie, 1994; Scholz & Schuler, 1993). An obvious advancement criterion is salary growth (Tziner et al., 1993). Salary leve1 or current job grade (which correlates highly with salary leve1 in most organizations) are used as criterion of management success by, for example, Hinrichs (1978) and Mitchel(l975). It is also used in the present study.

The Management Progress Study has produced mixed *results* (Thomton & Byham, 1982) with respect to the variation of assessment center validity with time. For the sample of 207 college graduates recruited in 195 7- 1960 as management trainees for a telephone company, the validity of assessment result and leve1 actually achieved, has decreased from a maximum of .46 in the early years (personal communication by Howard, 1981, in T'homton & Byham, 1982, p.254) to .33 in the sixteenth year. So the trend in validity appears to be inverted-U shaped. However, for the non-college group (n=148) the validity only decreased from .46 to .40 indicating a flattening validity curve with decreasing gains in predictive power. Also Slivinski, Grant, Bourgeois & Pederson (1977) found mixed results with respect to the predictive validity over time of the overall assessment rating (OAR) for salary. In the meta-analysis by Gaugler et al. (1987), there was no significant relation between assessment center validities and the time at which criterion measures were taken. In contrast with this, Tziner et al. (1993) investigating the validity of a managerial assessment center for a yearly rating of potential for upper management, found that the validity of the OAR for potential for upper-level management ***decreased*** with time. Finally, an *increase* in the validity of the OAR for advancement criteria such as rank attained or salary growth was found in the longitudinal studies by Anstey (1977), Hinrichs (1978), McEvoy & Beatty (1989), Mitchel (1975), and Moses (1971; see Huck, 1977). Concluding, studies on long-term assessment center validity have produced equivocal results; in addition, all were conducted in North-American or British organizations (Feltham, 1988).

Mitchel(l975) proposed two explanations for the variability of validity with time. The first explanation concentrates on changes in critical work elements as a consequence of organizational and societal developments, the second on job changes during an individual's career.

The first explanation of varying long-term assessment center validity refers to societal changes. For instance, Tziner et al. (1993) attribute the decrease in validity they found to general, organization-independent changes over time in what is needed to be a successful performer. We were not able to study this explanation in the present research.

A second explanation for varying long-term assessment center validities, is that being successful on the advancement criterion requires mastery of different combinations of the dimensions measured by the assessment center at different stages in the career. For instance, the predictive power of academic knowledge and written communication ski11 could be limited to early career stages, whereas self-confidence and oral communication could be predictive at long-range. In this case, it is the criteria which are dynamic instead of the persons (Hulin et al., 1990). In other words, job success factors change with tenure, and therefore the validities of assessment center dimensions predicting those factors also change over time. Note that in this case correlations between repeated measurements of the advancement criterion are expected to be low.

The longitudinal studies into assessment center validity discussed above suggest that aggressiveness, self-confidence, and oral communication (Hinrichs, 1978), impact (Mitchel), stress resistance, and organization and planning (Tziner et al., 1993) become more important with career progression. Career success not only depends on effective job performance, but also on being selected for higher levels jobs by decision makers (Luthans, Rosenkrantz & Hennessey, 1985). The latter requires both being interpersonally effective in the sense of demonstrating appropriate communication and networking behaviors so that you are noticed by decision-makers and admitted to the arena where the 'tournament' (Cable & Murray, 1999) for higher job positions takes place, and having the creativity, ambition and firmness of a 'proactive personality' (Seibert, Crant & Kraimer, 1999) that identifies opportunities, and shows initiative and perseverance in solving new managerial tasks. In line with this, research into management development (see e.g. Fiedler & House, 1994; Hogan, Curphy & Hogan, 1994) suggests the following determinants of career advancement in management:
-interpersonal effectiveness (networking, human relations, oral communication),
-persuasiveness (dominance, firmness, resistance to stress, self-confidence, decisiveness),
-achievement (ambition, energy level, drive).
Corresponding dimensions are used in the present study. We expect that these assessment center dimensions will gain in importance, and other dimensions (e.g. pertaining to academic knowledge) will lose their (early) validity with longer tenure.

We will also compare the long-term predictive validity of the assessment center to such other assessment instruments as the mental test and the selection interview. Whereas some researchers found that with respect to validity, assessment centers generally outperform other assessment instruments (Klimoski & Strickland, 1977), possibly with the exception of intelligence tests (Ree, Earles & Teachout, 1994), others observed less 'added value' with respect to traditional measures such as the mental test (Schmidt & Hunter, 1998). Another possibility is that assessment centers and other measures such as personality tests, measure different domains and therefore contribute independently to prediction (Goffin, Rothstein & Johnson, 1996).

METHOD

Subjects

Data were collected from the privatised Netherlands Postal and Telecommunications Services 'KPN'. The present study involved 679 academic graduates from Dutch universities recruited for a career in management in the years 1989-1995 by KPN. Persons did not have any significant prior work experience. The academic background of the recruits was (business) economics (35.2%) business administration (23.3%) engineering, technics and computer science (22.5%) management science (10.9%) and miscellaneous (e.g. law, history, or social science; 8%). The mean age at start of employment is 24.7 years (SD=2.4).

Men are a little older than women (mean age of men is 24.9 years, SD=2.3 years; mean age of women is 24.2 years, SD=2.4) because of obligatory duty in military service, which takes about one year and usually takes place after university graduation and before the first job. The percentage of females finally recruited in the period 1989-1995, varied between 26% to 45% with a mean of 34%. This average is large compared both to the low percentage of females in the student groups of interest (15%) and to the group of females among the applicants (17%).

In this study we compare predictor data collected during the selection procedure with criterion data on annual salary growth in the period of 2-7 years after time of hire. Loss of subjects due to turnover, incomplete or lost data records was 16% in total (see Table 1), which is comparable to the 12% loss in the study by Tziner et al. (1993), and much less than such percentages as 38% (Bray et al., 1974), 36% (Hinrichs, 1978), 38% (McEvoy & Beatty, 1989) and 56% (Mitchel, 1975).

Description of selection procedure

Selection dimensions

The following genera1 dimensions were assessed:

-'***thinking***': intelligence, cognitive fimctioning in all its aspects (cf. Stemberg, 1985): analytical reasoning, problem solving capability, creativity, imagination;

-'***interpersonal effectiveness:*** socially oriented and capable, interpersonally sensitive and competent, being open towards others and being able to deal with others, 'extraversion' from the 'Big Five';

-'***firmness***': independent, strong, decisive, resistant, stamina, able to cope with stress, dominance, self-confidence;

-'***ambition***': involvement, achievement motivation, commitment, energy level, drive.

-'***operational competence***': planning and organizing, productive, effective, systematic.

These dimensions initially were derived from studies into the primary dimensions that were used for management selection by (Anglo-)Dutch multinational companies such as Philips (Tigchelaar, 1974), Unilever and Shell (Muller, 1970; Ouwerschuur, 1988). But, in addition, they agree well with findings from research (as discussed above) into determinants of career advancement in management. And moreover, they correspond closely with results from research into the basic dimensions that, generally, underlie assessment center ratings (Sagie & Magnezy, 1997; Scholz & Schuler, 1993; Shore, Thomton & McFarlane Shore, 1990). In every step of the selection procedure, these rather broad dimensional categories are elucidated by corresponding behavioral samples or 'anchors' of the dimensions.

The selection procedure

The total selection procedure consists of the following consecutive and selective steps:

a. Selection based on the applicant's letter, by application of forma1 criteria as for instance field of study.
b. An interview conducted by a 'recruitment officer' (the 'recruitment interview'),
c. A mental test,
d. An interview with the manager who is in charge of the department where the candidate will start his or her career (the 'management interview'),
e. An assessment center, consisting of:
   el. A group discussion;
   e2. An analysis/presentation exercise,
   end meeting.

We take el and e2 as one step since in the fínal assessment center candidates can only be rejected on account of their overall assessment end rating (OAR) based on the combination of el and e2 (obtained in the end meeting).

Candidates start the procedure with step a, if accepted at step a then proceed to step b, if accepted at step b then proceed to c, . . . and so on, and fínally end with step e. Candidates can be rejected at every step.

Even when the end result of a step tums out to be positive, candidates can (and in fact do, see Table 1 below) withdraw from the procedure. After every step the candidate gets immediate feedback about his/her admission to the next step. But, knowledge obtained about a candidate at a previous step is never transmitted to the assessors who participate at a later step. In this study we concentrate on steps b-e.

Step b: Recruitment interview

'Recruitment officers' are members of the corporate recruiting department of the company. They have been trained in conducting selection interviews. The recruitment officer rates the candidate on the five dimensions, and also gives an overall rating. Next he/she decides on the continuance of the selection procedure with the candidate (which translating the overall 5-point rating to a yes/no rating; there was however no fixed, 'mechanical; rule for this).

Step c: Mental test

The *mental test* consisted of nine paper and pencil tests. Three have been developed within the company. The remainder are well-established and well-researched standardized intelligence tests which have quality high ratings in the Netherlands Test Documentation and Guidance Manual which is published and regularly updated by the Netherlands Association of Psychologists (Evers et al., 1992).

The tests result in one overall final score, and four so-called 'factor scores': on *numerical ability, analytical ability* (genera1 reasoning), *verbal ability,* and *creativity* ('divergent production' in the sense of Guilford, 1967: velocity and productivity of verba1 association). A cutoff score was determined for the final score. Persons with a fmal test score below the cutoff score were rejected.

Step d: Management interview

In the management interview the senior manager who is the 'owner' of the vacancy and consequently may become the manager of the graduate, decides on both the fit of the candidate for the job at issue, and hisiher potential for management development. Assessment dimensions and rating procedure were the same as in *the recruitment interview.* Only a fixed group of about 80 senior managers conduct the *management interview.* The same pool of senior managers participates, as assessor, in the final step of the procedure, the assessment center.

Step e: Assessment center

The final part of the procedure is an assessment center consisting of two situational exercises followed by an end meeting in which the OAR is determined. Up to this stage, candidates do not have contact with other candidates. However in the final assessment center, at least 5, and maximally 6 candidates participate together. They are, however, not competing with each other for jobs. The actual mix of the (5 or 6) persons partaking in a specific assessment center is determined only by the fact that each of those graduates has passed the preceding selection steps successfully.

For each center six assessors (or 5 depending on the number of candidates) are selected randomly from the pool of about 80 senior managers (the number of senior managers in this pool varied a little during the time period studied on account of organizational restructurings).

There is no relationship between assessor and candidate. In no case can a senior manager become the assessor of his 'own' candidate that is from a candidate he/she already has interviewed in the previous step of the *management interview.* In the (rare) case that a senior manager/assessor would have to assess during the assessment center the same candidate as observed by him/her before during the management interview, that senior manager was replaced by another senior manager selected randomly from the pool of assessors. Inexperienced senior managers are trained for the assessor% task by extensive and personal briefings beforehand, for instance by using samples of assessment center behavior on video.

Two exercises were used to elicit behavior from the candidates during the assessment center: a group discussion and an analysis/presentation exercise. The exercises and the end meeting take place on the same day. First six assessors observe and rate the behavior of 6 candidates in the *group discussion,* in which candidates have to come up with a common solution to a fictitious business problem in which they have conflicting interests. During the group discussion, all (6) assessors are present. In the analysis/presentation exercise the candidate has to present an analysis, and a corresponding plan of action, for another business problem. Only two assessors observe and discuss with the candidate his/her proposals and corresponding argurnents. In both exercises, assessors observe and rate the behavior of each candidate individually and separately on the dimensions *thinking, interpersonal effectiveness* and *firmness*. They do not give an end rating for the exercise and do not discuss their ratings before the end meeting.

The dimension ratings given by the individual assessors in the group discussion (18 ratings in total: 6 managers rate the candidate's behavior on 3 dimensions) and the analysis/presentation exercise (6 ratings in total: 2 managers rate the candidate's behavior on 3 dimensions) have to be combined to one final OAR. In the end meeting, candidates are discussed one after the other. For each candidate, managers start with reporting their observations and evaluations of his/her performance in the group discussion. Evaluations are compared and, in case of too large a difference, discussed in terms of underlying observations. The initial assessors' ratings of the dimensions, however, are not modifïed; they only serve as input for the discussion. When assessors feel that they have a clear view of the candidate on account of his or her behavior in the group discussion, they switch to the analysis/presentation exercise. In this case only two managers are able to bring in observations. The other managers can (and generally do) ask for clarifícation. On account of this new information, the 'picture' of the candidate is completed.

When assessors feel they have all the required information, they individually make their final decision. Every assessor gives his/her final rating of the candidate, based on both the own ratings from the group discussion and analysis/presentation exercises, and the discussion during the end meeting. Each assessor rates the candidate as 'insufficient' *(not acceptable, reject),* 'sufficient' (*average growth expectation, suited for the job but presently not a potential for top management),* or 'good' *(high growth expectation, potential top level manager).* The difference between 'sufficient' or 'good' indicates whether in the expectation of the assessors the candidate just is acceptable, or is a clear 'potential'. The category with the most ratings is taken as the final OAR.

In case of ties, the lower OAR is taken. For instance, if 3 assessors prefer the OAR 'insuffcient' and the 3 remaining assessors prefer the OAR 'sufficient', the final OAR will become 'insufficient' and the candidate is rejected. As a consequence of this conservative procedure, the distribution of the OAR will shift somewhat to the left: on the average, the clinical OAR will be somewhat lower than the 'actuarial' OAR computed as the average of the assessors' final ratings. The effect of this is that there will be a somewhat larger restriction in range on this predictor. *The OAR is not communicated,* neither to the candidate nor to the manager who will become his/her 'boss'. Candidates with ratings 'sufficient or 'good'are invited to join the company.

Predictors and criterion

Predictors

At the end of the recruitment interview and the management interview, the 5 assessment dimensions *thinking, interpersonal effectiveness, firmness, ambition* and *operational competence, are* rated on a five-point scale, ranging from 1 ('poor'), via 2 ('insufficient'), 3 ('average'), 4 ('good') to 5 ('very good'). For the two assessment center exercises of group discussion and analysis/presentation exercise, only the first three dimensions *thinking, interpersonal effectiveness,* and *firmness* are rated.

The final overall assessment rating (the 'OAR') refers to a three-point-scale, ranging from 0 ('insufficient; reject'), via 1 ('sufficient'), to 2 ('good').
Raw scores on the 9 paper and pencil test of the mental test are firstly recomputed into so-called 'factor scores' on the four factors of *numerical ability, analytical ability, verba1 ability,* and *creativity.* Factor analyses showed that 9 tests indeed measures these four genera1 and relatively independent intelligence factors. The factor scores are, secondly, transformed into a stanine normal distribution (Guilford, 1965) using the test score distributions of all academic graduates who have been tested for the company in the past 10 years (including persons who took part in selection procedures for other, non-managerial jobs). These 'Guilford 9 stanines' are, thirdly, transfoxmed into a 5-point scale ranging from 1 ('low', stanine 1; percentiles 0-4), via 2 ('below average', stanines 2 and 3; percentiles 4-23), 3 ('average', stanines 4, 5 and 6; percentiles 23-77), 4 ('above average', stanines 8 and 9; percentiles 77-96), to 5 ('high', stanine 9; percentiles 96-100). The final test score is computed as the average of the latter 5-point scale scores across the four dimensions of mental ability. Persons with scores 1 and 2 were rejected. We were not able to obtain the original raw test scores for the present study.

We studied the validity of two kinds of predictors: separate dimension (or factor) ratings as obtained in the recruitment interview, mental test, management interview, group discussion, and analysis/presentation exercise, and final ratings from the recruitment interview, mental test (final test score), management interview and assessment center (the OAR based on both group discussion and analysis/presentation exercise). Gaugler et al. (1987) could not obtain a reasonable estimate of the distribution of reliabilities of the OAR. Therefore in their meta-study this predictor was not corrected for reliability. In our study, the only predictor for which easily reliabilities can be obtained, is the mental test. In order to compute the reliabilities of the other selection steps, it is necessary to know which specific recruitment officer or manager/assessor participated in a selection step. This information was however not registered. Therefore we decided not apply corrections for reliabilities of the predictors.

Criterion

Career success was measured as average salary growth. Salary data were collected from November 1989 to November 1997. In the company investigated, salary leve1 is determined by:
-job grade: salary leve1 as determined by the position of the present job in the salary grading system. By this system jobs are weighted according to task load, knowledge and abilities, and responsibility.
-collective annual salary increases as a consequence of collective bargaining agreements between labor unions and the company
-individual merit increases on account of yearly appraisals of job performance.
In the period investigated, bargaining agreements resulted in collective annual salary increases of 2½% on the average. Consequently, managers with the same tenure but who started at different years will show different average salary growth figures only on account of these collective increases. For example, suppose the salary leve1 in 1989 is 100; then, taking only the effect of the collective yearly increases of 2½% into account, salary will be 107.7 in 1992, and average salary growth is 2.57. But for a manager with the same tenure of 3 years but who started his/her career in 1992, salary will increase from 107.7 (1992) to 116.0 in (1995), resulting in an average salary growth of 2.77. To avoid such an artificial difference in average salary growth, all salaries were corrected for collective increases. The difference between (corrected) last salary (obtained in November 1997) and (corrected) first salary was divided by the number of years the candidate had been working in the company. Following Gaugler et al. (1987) the reliability of the criterion of salary growth was assumed to be 1, that is we wished to be conservative and did not apply a correction for unreliability of the criterion.

RESULTS

First, we present a number of genera1 descriptive data with respect to predictors and the criterion. Next, we investigate the relationships between predictors and the salary growth criterion.

Preliminary analyses

Predictors

In Table 1 we present acceptation and rejection rates for all steps of the selection procedure. For example: In the period investigated, recruitment officers interviewed 4461 persons; 2302 of them (52%) were rejected. From the remaining 2159 persons with a positive end rating in the recruitment interview, 89 preferred to withdraw from the procedure leaving 2070 for the next step, the mental test.

Table 1.
Number of persons taking part in the selection procedure, rejected, accepted but withdrawing, and hired during 1989-1997

selection step	Total	Rejected	accepted but withdrawn	unknown
recruitment interview	4461	2302	89	
mental test	2070	241	178	
management interview	1651	222	89	
assessment center	1340	422	64	
salary negotiation	854		138	37
hired	679		126*	

*A total of 126 persons have left the company in the period investigated: 20 persons after 1 year, 29 after 2 years, 34 after 3 years, 21 after 4 years, 15 after 5 years, and 20 after 6 years.

Rejection percentages are 52% for the recruitment interview, 12% for the mental test, 13% for the management interview, and 3 1% for the assessment center. A mere 4% of the initial job applicants finally is hired. Table 1 shows that 679 persons finally were hired and therefore are part of the present study; 126 of them left the company somewhere during the period investigated.

For those persons for whom criterion data were available, Table 2 presents the number of persons, means and standard deviations of the ratings given on the dimensions assessed in the steps of the selection procedure. Since the assessors did not give final ratings for the group discussion and the analysis/presentation exercise, Table 2 gives an actuarial end score computed as the mean of all the dimension ratings given by the assessors in the exercise.

Table 2 also presents number of persons, averages and standard deviations for the total group of candidates, including those persons who were rejected somewhere in the selection procedure or who choose to withdraw from the procedure. In that way it is possible to estimate the degree of restriction in range in the predictors. Numbers in Table 2 vary on account of missing or incomplete data from 606 (mean of assessor ratings in the group discussion) to 679 (OAR).

Table 2.
Mean, and standard deviation for thepredictors for those managers for whom criterion data were available

Predictors	n	M	SD
RECRUITMENT INTERVIEW			
Thinking (*)	635 (3917)	4.09 (3.56)	.43 (.73)
Interpersonal effectiveness (*)	635 (3916)	4.09 (3.42)	.46 (.82)
Firmness (*)	634 (3911)	4.15 (3.52)	.42 (.80)
Ambition (*)	634 (3900)	4.10 (3.43)	.43 (.85)
Operational Competence (*)	634 (3870)	4.22 (3.69)	.40 (73)
Final rating (*)	641 (3921)	4.13 (3.46)	.32 (.74)
MENTAL TEST			
numerical ability (**)	633 (1725)	3.29 (3.22)	.74 (.80)
analytical ability (* *)	633 (1726)	3.28 (3.24)	.77 (.81)
verba1 ability (**)	633 (1724)	2.93 (3.00)	.72 (.78)
creativity (**)	633 (1690)	3.53 (3.29)	.76 (.79)
Final test score (**)	622 (1698)	3.34 (3.19)	.67 (.77)
MANAGEMENT INTERVIEW			
Thinking (*)	635 (1464)	3.98 (3.80)	.65 (.76)
Interpersonal effectiveness (*)	626 (1447)	3.95 (3.76)	.69 (.80)
Firmness (*)	634 (1461)	3.95 (3.76)	.67 (.78)
Ambition (*)	634 (1457)	3.82 (3.68)	.73 (.81)
Operational Competence (*)	629 (1430)	3.87 (3.72)	.72 (.79)
Final rating (*)	642 (1477)	3.95 (3.77)	.52 (.68)

GROUP DISCUSSION			
Thinking (*)	607 (1173)	3.76 (3.55)	.48 (.59)
Interpersonal effectiveness (*)	609 (1178)	3.60 (3.35)	.60 (.70)
Firmness (*)	608 (1175)	3.73 (3.50)	.58 (.70)
mean of assessors (*)	606 (1214)	3.74 (3.46)	.49 (.92)
ANALYSIS/PRESENTATION EXERCISE			
Thinking (*)	622 (1204)	3.67 (3.28)	.86 (1.04)
Interpersonal effectiveness (*)	620 (1200)	3.78 (3.41)	.79 (.95)
Firmness (*)	622 (1204)	3.94 (3.56)	.81 (1.00)
mean of assessors (*)	629 (1245)	3.83 (3.43)	.72 (.70)
Assessment Center FINAL RATING			
(clinical) Overall Assessment Rating (OAR) (***)	679 (1316)	1.38 (.93)	.49 (.75)

Note. Numbers, means and standard deviations for the total group, including persons who were rejected during the selection procedure and for whom therefore no criterion data were available, are reported in parentheses

*: Score range: 1 ('poor'), 2 ('insufficient'), 3 ('average'), 4 ('good'), 5 ('very good').

**: Score range: 1 ('low', stanine 1; percentiles 0-4), 2 ('below average', stanines 2 and 3; percentiles 4-23), 3 ('average', stanines 4, 5 and 6; percentiles 23-77), 4 ('above average', stanines 8 and 9; percentiles 77-96), to 5 ('high', stanine 9; percentiles 96-100).

***: Score range: 0 ('insufficient'; reject), 1('sufficient'), 2 ('good').

Table 2 clearly shows restriction of range effects. The average ratio of the standard deviations in the selected group to the standard deviations in the unselected group is .53 for the recruitment interview, .93 for the mental test, .86 for the management interview, .76 for the group discussion, .88 for the analysis/presentation exercise, and .65 for the OAR. There is a severe range restriction for the mean of the group discussion ratings (ratio is .53), but that there is no restriction in range for the mean of the ratings from the analysis/presentation exercise (ratio is 1.03). This suggests that in the assessment center end meeting the group discussion had more weight than the analysis/presentation exercise.

To check this, we regressed the OAR on the average rating of the group discussion and the ana.lysis/presentation exercise. The weight of the group discussion in predicting the OAR appeared twice the weight of the analysis presentation exercise.

weight of the analysis presentation exercise.

A (small) part of the restiction in range for the final OAR (ratio is .65) is caused by the conservative procedure to deal with ties among the assessors' final ratings. In 5% ofthe cases, there was a tie between 3 assessors prefening the OAR 'insufficient' and 3 assessors prefening the OAR 'sufficient'. In 7% of the cases, 3 assessors preferred the OAR 'sufficient' and 3 assessors the OAR 'good'.

It is conceivable that range restriction varies across the years because the selection ratio varies. For instance, for some reason recruitment officers could become more lenient in time, causing range restriction for the recruitment interview to decrease. In order to control for this, we computed selection ratios for every selection step and for every year. In the average, 4% of the total number of job applicants was hired every year. This selection ratio for the total selection procedure did not differ across the years included in the present study. There were only some minor variations in the selection ratios for the separate selection steps.

We also investigated whether there was a relation between tumover and ratings obtained in the selection procedure. In the group of persons with tenure of less than 2 years there were no differences between persons who left the company and persons who stayed. In the group with tenure between 2 and 5 years, there was a weak but significant positive correlation between tumover and ratings obtained on the dimensions *thinking* and *operational competence* in the recruitment interview (correlation in both cases was .08; $p<.05$), and the dimension *numerical ability* in the mental test (correlation is .13; $p< .01$). This implies that persons with better ratings in the recruitment interview and on one factor of the mental test are somewhat underrepresented when investigating long term validity. Also, there was a weak but significant negative correlation between tumover and ratings obtained on the dimensions *interpersonal effectiveness* (correlation is -.12; $p<.01$) and *firmness* (correlation is -.11; $p<.01$) in the analysis/presentation exercise, implying that there was a somewhat larger tendency to leave among persons with lower ratings in that assessment center exercise. These findings indicate that there is a smal1 additional restriction in range causing some underestimation of the long-term predictive validity of in particular the recruitment interview and the analysis/presentation exercise. In table 3, we present, for the group of selected persons, correlations between all predictors.

Table 3.

Correlations (computed on the group of selected candidates) between the dimensions thinking, interpersonal effectiveness, firmness, ambition and operational comwetence, and the end ratings (clinical final ratings or actuarial mean ratings) as assessed in the recruitment interview, mental test, management interview, assessment center group discussion, assessment center analysis/presentation exercise.

Selection step	Dimension	1 Recruitment interview a	b	c	d	e	f	2 Mental test a	b	c	d	e	3 Management interview a	b	c	d	e	f	4 Group discussion a	b	c	d	5 Analysis / presentation exercise a	b	c	d	6
1. Recruitment interview	a. Thinking																										
	b. Interpersonal Effectiveness	.17**																									
	c. Firmness	.14**	.38**																								
	d. Arnbition	.33**	.38**	.42**																							
	e. Operational Competence	.32**	.38**	.37**	.49**																						
	f. Final rating	.49**	.57**	.55**	.65**	.60**																					
2. Mental test	a. Numerical Ability	.07	-.02	-.01	.03	.01	.06																				
	b. Analytical ability	.02	-.02	-.06	-.12**	-.04	-.02	.40**																			
	c. Verbal ability	.oa	-.06	-.09*	-.oa	-.07	-.05	.17**	.24**																		
	d. Creativity	.06	.07	.05	.05	.05	.03	.11**	.15**	.10**																	
	e. Final score	.10*	.02	-.02	.01	-.oo	.05	.56**	.61**	.45**	.51**																
3. Management interview	a. Thinking	.05	.04	-.06	-.02	.04	.02	.01	-.03	.01	.01	-.01															
	b. Interpersonal Effectiveness	-.05	.10*	.03	.02	.05	.05	-.02	-.05	-.00	-.02	-.04	.34**														
	c. Firmness	-.07	.04	.05	-.02	.04	.01	-.06	-.05	-.07	.02	-.07	.41**	.38**													
	d. Ambition	-.05	.10	-.04	.06	.07	.02	-.03	-.05	.08*	-.05	-.06	.39**	.36**	.40**												
	e. Operational Competence	-.03	.06	.01	.01	.09	.07	-.oo	-.07	.07	-.03	-.II	.39**	.43**	.43**	.38**											
	f. Final rating	-.04	.06	.oo	.02	.06	.04	-.03	-.06	-.05	-.04	-.09	.63**	.66**	.65**	.63**	.65**										
4. Group discussion	a. Thinking	.06	.06	.01	.07	-.00	.04	.04	.02	.03	.01	.01	.Ol	.05	.03	.oo	.02	.oo									
	b. Interpersonal Effectiveness	-.02	.12**	.01	.04	-.05	.04	.00	.02	.oa	.oo	-.01	-.04	.10*	.04	.00	.04	.02	.63**								
	c. Finness	03	.12**	.02	.07	-.04	.04	.02	.02	.05	.02	.oo	.oo	.07	.09	.07	.03	.06	.66**	.72**							
	d. Mean rating	.00	.11**	.01	.06	-04	.04	.02	.02	.06	.03	.oo	-.01	.10	.07	.04	.05	.05	.85**	.89**	.90**						
5. Analysis / presentation exercise	a. Thinking	.09	0 1	-.01	.04	-.06	.03	.01	.04	.15**	.02	.07	.11**	-.02	-.02	-.03	.05	.05	.22**	.09*	.11**	.17**					
	b. Interpersonal Effectiveness	-.05	.06	-.01	-.01	-.09*	.01	-.04	-.01	.07	.oo	.01	.05	.12**	.07	.07	.05	.09	.24**	.34**	.28**	.34**	.51**				
	c. Finness	-.07	.00	.02	.01	-.10*	-.02	-.02	-.01	-.01	.07	.00	.06	.10*	.oa	.09*	.07	.10*	.25**	.25**	.29**	.30**	.54**	63"			
	d. Mean rating	-.00	.03	-.02	.05	-.07	.02	-.04	-.01	.03	.03	.02	.09	.08*	.06	.06	.oa	.10**	.27**	.24**	.25**	.29**	.77**	.79**	.81**		
6. Overall Assessment center end Rattng		.01	.08	-.00	.07	-04	.05	-.01	.02	.06	03	.02	.05	11"	.06	05	.05	.06	.66**	.70**	.69**	.78**	46"	.58**	.57**	64**	

Note N varies between 581 and 679. *p<.05. * p<.01.

Table 3 shows that the selection steps are independent. The group discussion and the analysis/presentation exercise are correlated. The large correlations between the OAR and the end scores of both assessment centers was to be expected. In accordance with previous studies (e.g. Brannick, Micheals & Baker, 1989; Klimoski & Brickner, 1987) the correlation between the same predictors across different assessment situations ('within dimension, between methods' correlation) generally is lower than the correlation between different predictors within the same situation ('between dimensions, within method' correlation). For *thinking the* average between-method correlation is .09, for ***interpersonal effectiveness*** it is .14, and for *firmness* .09. But, 5 of the 6 between-methods correlations for ***interpersonal effectiveness*** reach signitìcance.

The average within-method correlation is .23 for the recruitment interview, .38 for the management interview, .67 for the group discussion, and .56 for the analysis/presentation exercise. The latter two average correlations indicate that the manager-assessors did not discriminate very well between the dimensions in the two assessment center exercises. The average within-method for the assessment center is .62. Although this correlation is high, it is not untypical for assessment center studies. For instance, Bycio, Alvares & Hahn (1987) found an average within method correlation of .75, and Schneider & Schmitt (1992) of .72. The within-dimension correlations across the two assessment center exercises *are* .22 *(thinking),* .34 (***interpersonal effectiveness***), and .29 (*firmness*; all p< .001; see Table 4). The average within-dimension correlation is .28. For comparison, Bycio et al. (1987) obtained an average within-dimension correlation of .36, and Schneider et al. (1992) of .25.

In regard of these results, we will present validity results for dimensions per selection step below.

Criteria

The number of people that had a tenure of 8 years was too smal1 to use their salary data. Generally, turnover is higher for low performers (Trevor, Gerhart & Boudreau, 1997). Taking average salary growth as an indicator of overall job performance, it is expected that persons whose salary growth stayed low or decelerated had a greater tendency to leave. Indeed, the correlation between average salary growth (corrected for starting salary) and quitting the company over the period investigated is -.26 (N=605, p<.001). This implies that in particular at later time points, the standard deviation of the criterion will decrease, This was the case in our study. There is a dip in the standard deviation of the criterion at five years of tenure; however, from there it rapidly increases again. Therefore we expect that this kind of restriction in range on the criterion will not affect long-term validity.

Finally, we computed for the fixed group of 88 persons who were hired in 1989 or 1990, their average salary growth at 2, . . ., 6 years of tenure (in this case salary growth was not corrected for starting salary since we did not expect differences in starting salary would affect the criterion correlations; in fact, they did not). Then we correlated these 5 criterion measures. In accordance with the idea of dynamic criteria, correlations between average salary increases became lower when the number of years between the times of criterion measurement increased (cf. Guion, 1997). For instance, the correlation between average salary growth measured after 4 years and average salary growth measured after 5 years was .78 (p<.01; N=82). But, the correlation between average salary growth measured after 2 years and average salary growth measured after 7 years was .18 (p=. 10; N=86). In the average, the criterion intercorrelation decreased with about .15 for every one-year difference between the time points of criterion measurement. Note that when criteria are dynamic, different managers will excel at different time-points, which implies that predictor validities will also change in time.

Validity

In order to investigate time-patterns in assessment center validities, we first present an overview of validities for different tenure levels in Table 4. In order to compare the long-term assessment center validity to long-term validities of other selection instruments, we then investigate the interaction between selection steps, and the effect of tenure on validity in Table 5.

Table 4 gives the correlations between the predictors: all dimensions from the selection steps of the recruitment interview, mental test, management interview, group discussion and analysis/presentation exercise, and the criterion average salary growth for different tenure levels. The correlations are corrected for the effect of different starting salaries by partialling out the initial salary. Between parentheses we give correlations corrected for both starting salary and restriction in range. Note that we did not corrected for criterion unreliability.

Table 4

Partial correlation between predictors and average salary growth, with starting salary partialled out, for managers with 2, .., 7 years of tenure.

Predictors	Tenure in years (number of persons)					
	2 (562)	3 (524)	4 (408)	5 (311)	(219)	7 (88)
Thinking	.06 (.10)	.03 (.05)	.01 (.02)	.06 (.10)	.04 (.07)	-.06 (-.10)
Interpersonal Effectiveness	.09* (.16)	.05 (.09)	.02 (.04)	.10 (.18)	.09 (.16)	-.05 (-.09)
Firmness	.05 (.09)	-.09* (-.17)	.01 (.02)	.07 (.13)	.16* (.29)	.15 (.28)
Ambition	.09* (.18)	.03 (.06)	.03 (.06)	.19** (.36)	.14 (.27)	.05 (.10)
Operational competence	.06 (.11)	-.02 (-.04)	.02 (.04)	.10 (.18)	.03 (.15)	-.15 (-.27)
Final Rating	.06 (.14)	-.02 (-.05)	-.01 (-.02)	.12* (.27)	.12 (.27)	-.04 (.09)
Mental Test						
Numerical ability	.00 (.00)	-.06 (-.06)	.11* (.12)	.04 (.04)	.08 (.09)	.13 (.14)
Analytical ability	-.06 (-.06)	-.06 (-.06)	.02 (.02)	-.07 (-.07)	-.07 (-.07)	.00 (.00)
Verbal ability	-.08 (-.09)	-.10* (-.11)	-.04 (-.04)	-.19** (-.21)	-.22** (-.24)	-.18 (-.19)
Creativity	.05 (.05)	-.03 (-.03)	-.01 (-.01)	.03 (.03)	.03 (.03)	.03 (.03)
Final rating	-.01 (-.01)	-.07 (-.08)	.05 (.06)	-.04 (-.05)	-.05 (-.06)	-.01 (-.01)
Management Interview						
Thinking	.05 (.06)	.09 (.11)	.03 (.04)	-.02 (-.02)	-.00 (-.00)	-.06 (-.07)
Interpersonal Effectiveness	-.01 (-.01)	.10 (.12)	.04 (.05)	-.03 (-.03)	.03 (.03)	-.13 (-.15)
Firmness	.07 (.08)	.11* (.13)	.11* (.13)	.09 (.10)	.14* (.16)	.11 (.13)
Ambition	.02 (.02)	.13* (.14)	.06 (.07)	.08 (.09)	.12 (.13)	.01 (.01)
Operational competence	.11* (.12)	.10* (.11)	.02 (.02)	.09 (.10)	.13 (.14)	-.04 (-.04)
Final Rating	.02 (.03)	.13* (.17)	.10* (.13)	.07 (.09)	.11 (.14)	-.01 (-.01)

Assessment Center group discussion						
Thinking	.06	.01	-.02	.03	.14	.24
	(.07)	(.01)	(-.02)	(.04)	(.17)	(.29)
Interpersonal Effectiveness	.03	.04	.06	.16*	.21*	.23
	(.03)	(.05)	(.07)	(.19)	(.24)	(.27)
Firmness	.13**	.11*	.10	.16*	.18*	.30*
	(.16)	(.13)	(.12)	(.19)	(.22)	(.35)
Final Rating	.09*	.07	.06	.13*	.20*	.29*
	(.17)	(.13)	(.11)	(.24)	(.36)	(.49)
Assessment Center analysis/presentation exercise						
Thinking	.13**	.04	-.01	-.00	.03	.01
	(16)	(.05)	(-.01)	(-.00)	(.03)	(.01)
lnterpersonal Effectiveness	.08	.11*	.06	.08	.08	.17
	.04	(.13)	(.07)	(.10)	(.10)	(.20)
Firmness	(.05)	.05	.Ol	.05	.oo	.15
		(.06)	(.01)	(.06)	(.00)	(.18)
Final Rating	.10*	.08	.Ol	.03	.03	.07
	(.14)	(.11)	(.01)	(.04)	(.04)	(.10)
***Final** rating*						
(clinical) Overall assessment center rating	.13**	.08	.02	.05	.13	.27*
	(.20)	(.12)	(.03)	(.08)	(.20)	(.39)

Note. Between parentheses correlations corrected for restriction in range. We did not apply a correction for criterion unreliability
*p<.05. ** p< .01.

Table 4 shows that, generally, validities are moderately in size (although significant). In addition, validities vary with tenure. Only in a few cases are the validities consistently positive for all tenure levels. *Firmness* both as assessed in the management interview and the group discussion is such a consistent predictor throughout the whole period. *Verbal ability* (mental test) is a consistent negative predictor. For the other dimensions, it seems that the validity is limited to a part of *the* career. For *instance, interpersonal effectiveness* as assessed in the group discussion seems to be particularly predictive for higher tenure levels. The same can be observed for the final ratings of the exercises. The group discussion is a consistent positive predictor for all tenure levels. The management interview is predictive early in the career, and the OAR is predictive early and later in the career. For the other instruments there are no clear pattems.

The foregoing suggests that selection steps and corresponding dimensions are predictive at different stages of the career. Therefore, we investigated time-dependent changes in the relation between long-term assessment center validity and long-term validities of other selection instruments, as follows. The interaction between the predictive validity of dimensions within a selection step and tenure was tested with hierarchical moderated regression analysis (Cohen & Cohen, 1975). The dependent variable was total salary growth, computed as the difference between the salary leve1 in 1997 and the starting salary level. First, we included tenure as a control. Then, all predictors and the interactions between tenure and predictors were included. The final ratings of the selection steps were not included as predictors since they are correlated with the dimensions. Since observed validities are lowered by restriction in range, the threshold for incorporation in the regression equation was decreased. Instead of the customary 5% probability of including an additional predictor by which the multiple correlation increases on account of pure chance only, we opted for a 10% type I error probability; the assumption being that the real probability of including an additional predictor wrongly would hardly increase.

Table 5

Hierarchical moderated regression of total salary growth on (step 1) tenure (control) and (step 2) the selection steps and the interaction between tenure and selection step

Independent variable	Standardized coefficient beta*	Sign. of beta	Multiple R^2	F of multiple R
Constant (unstandardized)	1155	.43		
Step 1				
Tenure	.21	.45	.37	193.92*
Step 2.				
1. Interact. AC-GD: FI	.35	.00	.40	109.57*
2. Interact. MT: VA	-.77	.01	.43	82.47*
3. Interact. MT: NA	.67	.01	.44	63.94*
4. MT: CR	.097	.03	.45	52.31*
5. Interact. RI: A M	.241	.06	.45	44.56*
6. MT: NA	-.332	.03	.46	38.90*
7. MT: VA	.302	.09	.47	34.60*

Note. Total salary growth was computed as the difference between salary in 1997 and starting salary.
R = Multiple Correlation; RI: recruitment interview; MT: mental test; AC-GD: assessment center group discussion; FI: firmness; AM: ambition; NA: numerical ability; VA: verbal ability; CR: creativity.
Betas and probability values are computed on account of the final regression equation with all predictors. N=464; listwise deletion.
*p<.001
(To be published. Do not quote without permission of the authors.)

In Table 5 we present the results. We both present standardized betas and their p-values, and multiple R's and their p-values. The betas are based on the final equation obtained with all predictors. The multiple R2 for this final equation was .47 (see the last line in Table 5; p<.001; adjusted R2=.45). The pattern of beta's in Table 5 shows that significant predictors are *numerical ability, verba1 ability,* and *creativity* from the mental test. Significant moderator effects of tenure were found for the relation between (absolute) salary growth and *ambition* as assessed in the recruitment interview, *numerical ability* and *verba1 ability* from the mental test, and *firmness* as assessed in the group discussion. The regression analysis generally confirms the results from table 4. In addition, it appears that *ambition* (recruitment interview) and *numerical ability* (mental test) increase in validity, and that *creativity* (mental test) is a moderate predictor throughout the career. To check these findings we limited the analysis to those persons with tenure of at least 7 years, and again took average salary growth as criterion. We did not correct for restriction in range or criterion unreliability. In that case, after 2-3 years the only (almost) significant predictor is *thinking* as assessed in the recruitment interview (validity is .27; p<.08; N=88). After 6-7 years the only significant predictor is *firmness* as assessed in the group discussion (validity is .39; p<.01; N=88). The pattern after 4-5 years is much less clear.

Additional analyses

Since it was not registered which recruitment officer or manager participated in which interview, it was not possible to compute reliabilities for the recruitment interview and the management interview.
The four factors of the mental test all had high reliabilities; the mean reliability was .78. As a consequence, the predictive validity of *numerical ability* for persons with 7 years of tenure, corrected for restriction in range and for starting salary, increases slightly from .14 to .16. It was not registered which assessor took part in which assessment center. We did the following to obtain an estimate of the reliability of the dimension ratings of the group discussion. For each group discussion with 6 assessors, we computed Cronbach's alpha coefficient for the ratings of these assessors of the same dimension (these data comprised both selected persons and persons who were rejected later in the procedure). Next for every dimension we computed the mean and standard deviation of these coefficients across those group discussions where sufficient data were available and in which 6 assessors participated. The following average reliability estimates were obtained across 70 group discussions with a total of 420 candidates: .65 (*thinking*; SD= .26), .68 (*interpersonal effectiveness*; SD= .30), and .65 (*firmness*; SD= .37). Although the averages indicate acceptable reliability, the standard deviations indicate that the group discussions varied considerably with respect to the reliability of the dimension ratings. In some (rare) cases, the assessors did not agree at all with each other. Applying these values, the validity of *thinking* for persons with 7 years of tenure increases from .29 to .36, for *interpersonal effectiveness* from .27 to .33, and for *firmness* from .35 to .43. It was not possible to apply the same procedure to compute reliabilities for the analysis/presentation exercise.
An estimate of the reliability of the OAR can be obtained in the same way as the reliability of the group discussion dimensions was obtained. The average reliability estimate of the OAR obtained across 77 assessment centers with a total of 462 candidates is .77 (SD=. 19). Applying this value, the validity of the OAR for predicting average salary growth over a period of 7 years, corrected for initial differences in starting salaries, for restriction in range, and for unreliability of the predictor, becomes .44.

We also did some additional analysis to obtain more insight into the overall negative validity pattern of the factor *verbal ability* of the mental test. For instance we took gender as an extra control (females typically are somewhat higher on verba1 ability and males on numerical ability; this was also the case in the present sample), but the pattern remained. Table 2 shows that the average test score on *verbal ability* is relatively low for the selected group. In fact, the average for the selected group even is a little lower than the average for the total applicant group. To find out if by including a mental test, the company is inadvertently selecting in an adverse way on *verba1 ability*, we regressed the final test score on all test factors. It turned out that all factors contributed positively to the final test score, although the regression weight of *verba1 ability* was smaller than the weights of the 3 other test factors.

In this study we investigated the predictive value of the *clinical'* OAR. In addition, we computed an actuarial assessment center end rating. First, average dimension ratings are computed for the group discussion, and the analysis/presentation exercise, across all assessors. Then, an exercise end rating is computed as the average of the three average dimension ratings. Finally, the actuarial assessment center end rating is computed as the average of the end ratings of the two exercises. It appeared that the actuarial OAR was strongly correlated with the clinical OAR: .79 (N=1153; p< .OOl). As expected, therefore the validities, computed as the partial correlations between the actuarial OAR and average salary growth (with starting salary partialled out), for different tenure levels are almost the same as those for the clinical OAR. For 2, 3, . . ., 7 years of tenure the partial correlations are .11* (.13**), .07 (.08), .04 (.02), .10 (.05), .14* (.13), .26 (.27*) respectively (*:p<.05; ** :p<.0 1; between parentheses the corresponding validities for the clinical OAR obtained from the bottom line of Table 4). This again confirms that generally mechanica1 assessment center composite scores result in similar prediction (Petersen & Pritz, 1986).

DISCUSSION

The long-term validity of the OAR, corrected for initial differences in starting salaries and for restriction in range, is .39 after 7 years. This agrees with the mean validity for career advancement, corrected for statistical artifacts such as sample size and restriction in range, of .36 that Gaugler et al. (1987) obtained in their meta-study. For the same type of criterion, Bray et al. (1974) obtained a validity of .32 (N=123; 8 year period) for the dimension 'human relations', while Hinrichs (1978) found a validity of .40 (N=30; 8 year period) for the dimension 'interpersonal contact'. Taking as a predictor the average of the dimension ratings across group discussion and analysis/presentation exercise, the long-term validity of the corresponding dimension *interpersonal effectiveness* for total salary growth in our study becomes .26 (N=83; p=.06). In the same way: for the dimension resistance to stress, Bray et al. found a validity of .3 1; Hinrichs also found .3 1, whereas we obtained a long-term validity of .32 (p=.01) for the corresponding dimension *firmness*. So, in the long run, the assessment center is a good predictor of such dimensions as interpersonal effectiveness and firmness. The cognitive dimension *thinking* was not predictive at any moment.

There was however a considerable time variation in the validity. The OAR predicts average salary growth in the first years and in the final years. In between, that is for persons with 3-5 years of tenure, the OAR is not related to career advancement. There was a corresponding time variation of the dimensions: When we add the mental test and both interviews as selection steps, things becomes complicated because dimensions were, although not unexpectedly, only marginally consistent across steps. However, it appeared that every step contributed unique dimensions to the prediction. The group discussion contributes with *interpersonal effectiveness* and in particular *firmness*. The interviews, in particular the recruitment interview and to a minor degree the management interview, contribute *with ambition* and predict career progression in later years. The mental test contributes the cognitive dimensions of *numerical ability* and *creativity;* later in the career persons higher on *numerical ability* make faster career progression, whereas *creativity* is predictive during the whole period. Most of this pattem was as expected on account of studies in managerial effectiveness and development: persuasive and decided behaviors are a constant determinant of management progress, while the impact of interpersonal and achievement behaviors gradually increases.

Since construct-validity of the dimensions was low, we can, for our interpretation of the results, as well switch from a person-based (dimensions) to a task-based (exercises or selection steps) interpretation (Russel & Domm, 1995). This is reinforced by the fact that, since in our study every step only contributed unique dimensions to the prediction, dimensions and exercises more or less merge. Note that from a predictive point of view, this is no problem. Studies (see e.g. Jones et al., 1991; Sackett & Wilson, 1982) indicate that it does not matter what the OAR indicates: a person-based dimensional profile or an exercise-based situational profile, as long as it is only predictive validity that counts. The OAR is valid because it stands for a large sample of behavioral evidences.

In a task-based interpretation, there is a close correspondence between assessment tasks and critical task domains one has to master successively when developing as a manager. Research by McCauley, Ruderman, Ohlott & Morrow (1994) shows that critical for management development are such 'developmental job components' as dealing with unfamiliar responsibilities, developing new directions, solving problems with employees, handling job overload and extemal pressure, and influencing without authority. All these very closely resemble tasks in the group discussion. Career development consists of the successive mastery of such job components. It is not a smooth, continuous process but consists of steep stages and thresholds corresponding for instance of having to deal with people management tasks or commercials activities for the first time. Individual variations in the first occurrence of such critical job components may account for the less clear validity patterns at intermediate tenure levels.

Schmidt and Hunter (1998) present an overview of the predictive validity of a range of selection instruments for the criterion of overall job performance. In addition, they determine the incremental validity of these predictors with respect to measures of genera1 mental ability. They found that the gain from adding an assessment center to a mental test is low since there generally will be a large correlation (.50 on the average according to Schmidt & Hunter) between the two instruments.

The gain from adding a structured interview (mean correlation with assessment center is .30) is estimated to be much larger. However, in our study the correlations between the end ratings of the recruitment interview, mental test, management interview, group discussion, and analysis/presentation exercise were much lower than these (see Table 3). There are two reasons for this.
First, the group that started the selection procedure already is restricted with respect to intelligence and ability on account of self-selection. Second, on account of the hierarchical set up of the selection procedure those persons who make it to the assessment center, are already 'homogenized' with respect to their scores on both interviews and the mental test. As a consequence, predictors have low intercorrelations, and there is room for the contribution of all instruments (selection steps) to the prediction. Third, it is conceivable that the gain from adding an assessment center to for instance a mental test varies with the time-point of criterion measurement. If validities are dynamic, it is likely, both from a psychological and a statistical point of view, that incremental validities are dynamic too.

*Limitations **of** the present study and implications forfurther studies*
An explanation for the negative validity of *verba1 ability* is given both by nature of the organization where we did our study, and by the fact that the selection procedure was severe. The organization was a postal and telecommunications company with a technical core, which moreover was recently privatized. In such a domain specific intelligence factors, such as *numerical ability* and *creativity* may become more important than *verba1 ability*. Moreover, in accordance with the existence of a genera1 intelligence factor g, the four factors of the mental test had fairly large intercorrelations in the total group of applicants. The average correlation was .47 (N=1721). After the selection procedure however, the average correlation dropped to .28 (N=644), which is a genera1 finding when correlations are computed on higher cognitive ability groups. It is uncommon that persons score rather high (or rather low as the case may be) on all test factors in selected groups. Legree, Pifer and Grafton (1996, p.55) illustrate this by the statement that *"Albert Einstein might have been a mediocre historian"*. The average score on verba1 ability indeed was considerably lower than the average of the other test factors. Both findings imply that there is little room for ***verba1 ability*** as a predictor. By the artificial limitation of the selected group to the top of the distribution of genera1 intelligence, the correlation between verba1 ability and the other test factors is low, and the correlation with the criterion becomes even negative given the nature of the company.

We found low correlations between the same dimension assessed in different steps. One reason for this is that, by the hierarchical design of the procedure, the selection steps of interview(s), test and assessment center are relatively independent. Another, that the steps differed in type of assessors with respect to functional background, amount of training, and experience in assessment. Recruiters in the recruitment interview had more experience and more training (some of them were professional psychologists) than managers in the management interview, assessment center group discussion and assessment center analysis/presentation exercise. Presence of professional psychologist as assessors is a robust moderator of assessment center validity (Gaugler et al. 1987; Sagie & Magnezy, 1997).

In this study, the analysis/presentation exercise case had low predictive value. A possible explanation is that the reliability of the end score of the group discussion is larger than the reliability of the analysis/presentation exercise. The group discussion is based on the average of 5 or 6 assessors, whereas the end rating of the analysis/presentation exercise is based on only 2 assessors. Unfortunately, we were not able to compute the reliability of the latter exercise.

In conclusion, conflicting findings with respect to long-term assessment center validity can be explained by changes in determinants of job success, which are related to changes in job demands when a person advances in his/her career. By investigating such changes we can explain the differences and dynamics in validities of the assessment center and other selection instruments such as an employment interview and a test for genera1 mental ability.

REFERENCES

Anstey, A. (1977). A 30-year follow-up of the CSSB procedure, with lessons for the future. Journal of Occunational Psychology, 50, 149-159.

Barrett, G.V., R.A. Alexander, & D. Doverspike. (1992). The implications for personnel selection of apparent declines in predictive validities with time: A critique of Hulin, Henry, and Noon. Personnel Psychology, 45, 601-617.

Brannick, M.T., Micheals, Ch.E., & Baker, D.P. (1989). Construct validity of in-basket scores. Journal of Applied Psychology, 74, 957-963.

Bray, D. W., Campbell, R.J., & Grant, D.L. (1974). Formative vears in business: A long-term A.T.&T. studv of managerial lives. New York: Wiley.

Bycio, P., Alvares, K.M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. Journal of Applied Psychology, 72, 463-474.

Cable, D.N., & Murray, B.N. (1999). Toumaments versus sponsored mobility as determinants of job search success. Academy of Management Journal, 42 (No.4), 439449.

Cohen, J. & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral science. Hillsdale, New Jersey: Lawrence Erlbaum

Evers, A., Vliet-Mulder, J.C., van, & Laak, J., ter (1992). Documentatie van tests en testresearch in Nederland (Documentation of tests and test research in the Netherlands). Nederlands Instituut van Psychologen. Assen: Van Gorcum.

Feltham, R. (1988). Validity of a police assessment centre: A 1- 19 year follow up, Journal of Occupational Psychology, 61, 129- 144.

Fiedler, F.E. & House, R.J. (1994). Leadership theory and research: A report of progress. In: C.L. Cooper & I.T. Robertson (Eds.). Key reviews in manaeerial psychology. New York: Wiley, 97-116.

Gaugler, B.B., Rosenthal, D.B., Thomton 111, G.C., & Bentson, C. (1987). Meta-analysis of assessment centre validity. Journal of Applied Psychology, 72, 493-5 ll.

Goffin, R.D., Rothstein, M.G., & Johnston, N.G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. Journal of Applied Psychology, 81 (No.6), 746-756.

Guilford, J.P. (1965). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Guilford, J.P. (1967). The nature of human intelligence. New York.

Guion, R.M. (1997). Criterion measures and the criterion dilemma. In: N. Anderson & P. Herriot (Eds.). International Handbook of Selection and Assessment. New York Wiley, 267-286.

Hinrichs, J.R. (1978). An eight-year follow-up of a management assessment center. Journal of Applied Psychology, 63 (No.5), 596-601.

Hogan, R., Curphy, G.J., & Hogan, J. (1994). Wat we know about leadership. Effectiveness and personality. American Psychologist, 49 (No.6, June), 493-504.

Huck, J.R. (1977). The research base. In: J.L. Moses & W.C. Byham (Eds.). Applying the assessment center method. New York: Pergamon Press, 26 1-29 1.

Hulin, Ch.L., R.A. Henry, & S.L. Noon. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. Psychological Bulletin, 107 (No.3), 328-340.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of some altemative predictors of job performance. Psychological Bulletin, 96, 72-98.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment centre, Journal of Occunational Psychology, 64, 1-2 1.

Klimoski, R., & Brickner, M. (1987). Why do assessment centres work? The puzzle of assessment centre validity. Personnel Psychology, 40, 243-260.

Klimoski, R., & Strickland, W.J. (1977). Assessment centers: Valid or merely prescient? Personnel Journal, 30, 353-361.

Legree, P.J., Pifer. M.E., & Grafton, F.C. (1996). Correlations among cognitive abilities are lower for higher ability groups. Intelligence, 23, 45-57.

Luthans, F., Rosenkrantz, S.A., & Hennessey, H.W. (1985). What do successful managers really do? An observational study of managerial activities. Journal of Applied Behavioral Science, 21 (No.3, Aug), 255-270.

McCauley, C.D., Ruderman, M.N., Ohlott, P.J., & Morrow, J.E. (1994). Assessing the developmental components of managerial jobs. Journal of Applied Psychology, 79, 544-560.

McEvoy, G.M., & Beatty, R.W. (1989). Assessment centres and subordinate appraisal of managers: A seven year examination of predictive validity. Personnel Psychology, 42, 37-52.

Mitchel, J.O. (1975). Assessment center validity: A longitudinal study. Journal of Applied Psychology, 60 (No.3), 573-579.

Moses, J.L. (1971). Assessment center performance and management progress. Paper presented at the 79" annual meeting of the American Psychological Society, Washington, D.C..

Muller, H. (1970). The search for the qualities essential for advancement in a large industrial group (Shell). Doctoral dissertation. Utrecht, The Netherlands: University of Utrecht.

Ouwerschuur, F.K.G. (1988). Investeren in managers (Investing in managers). Bestuursjournaal, nr.1 (februari), 17-20.

Ree, M.J., Earles, J.A., & Teachout, M.S. (1994). Predicting job performance: Not much more than g. Journal of Applied Psychology , 79, 5 18-524.

Peterson, D.K., & Pitz, G.F. (1986). Effect of input from a mechanica1 model on clinical judgment. Journal of Applied Psychology, 71, 163- 167.

Ritchie, R. J. (1994). Using the assessment center method to predict senior management potential. Special issue: Issues in the assessment of management and executive leadership. Consulting Psychology Journal; Practice & Research, 46 (No. 1), 16-23.

Russell, C.J., & Domm, D.R. (1995). Two field tests of assessment centre validity. Journal of Occupational and Organizational Psychology, 68, 25-47.

Sackett, P.R., & Wilson, M.A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. Journal of Applied Psychology, 67, 10-17.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of indistinguishable dimension categories, and assessment centre construct validity. Journal of Occupational and Organizationa 1 Psychology, 70, 103-108.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practica1 and theoretical implications of 85 years of research findings. Psychological Bulletin, 124 (No.2), 262-274.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Schneider, J.R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. Journal of Applied Psychology, 77, 32-41.

Scholz, G., & Schuler, H. (1993). Das nomologische Netzwerk des Assessment centers: Eine Meta-analyse (The nomological network of the assessment center: A meta-analysis). Zeitschrift fur Arbeits- und Organisationspsychologie, 37 (No.2), 73-85.

Seibert, S.E., Crant, J.M., & Kraimer, M.L.. (1999). Proactive personal@ and career success. Journal of Applied Psychology, 84 (No.3), 416-427.

Shore, T.H., Thomton, G.C. III, & McFarlane Shore, L. (1990). Construct validity of two categories of assessment center dimension ratings. Personnel Psychology, 43, 10 1 ▪ 116.

Slinvinski, Grant, Bourgeois & Pederson (1977). Development and application of a first leve1 assessment centre. Ottawa, Canada: Managerial Assessment and Research Division of the Personnel Psychology Centre.

Stemberg, R.J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge: Cambridge University Press.

Thomton, G.C. III, & Byham, W.C. (1982). Assessment centers and managerial performance. New York: Academic Press.

Tigchelaar, L.S., (1974). Potentieel beoordeling en loopbaansucces (Potential appraisal and career success). Akademisch Proefschrift. Universiteit van Amsterdam.

Trevor, C.O., Gerhart, B., & Boudreau, J.W. (1997). Voluntary tumover and job performance: Curvilinearity and the moderating influence of salary growth and promotions. Journal of Applied Psycholoq, 82 (No. 1), 44-61.

Tumage, J.J., & Muchinsky, P.M. (1982). Transsituational variability in performance within assessment centres. Organizational Behavior and Human Performance, 30, 274-300.

Tziner, A., Ronen, S., & Hacohen, D. (1993). A four-year validation study of an assessment center in a financial corporation. Journal of Organizational Behavior, 14, 225-237.

Authors Note

Paul G.W. Jansen, Department of Management & Organization Studies, Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam, the Netherlands;
Bert Stoop, former Researcher at KPN Research, Groningen, The Netherlands.

We thank Dan Simunic (Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada), Marke Bom (Department of I/O Psychology, Vrije Universiteit Amsterdam), and Mandy van der Velde (Department of Management & Organization Studies, Vrije Universiteit Amsterdam), for their helpful comments on earlier versions of this paper.

Correspondence conceming this article should be addressed to Paul G.W. Jansen, Department of Management & Organization Studies, Faculty of Economics and Business Administration, Vrije Universiteit,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
Electronic mail may be sent to pjansen@feweb.vu.nl